

# Structure-Preserving Iterations on Matrix Groups

Nick Higham  
Department of Mathematics  
University of Manchester

higham@ma.man.ac.uk  
<http://www.ma.man.ac.uk/~higham/>

Joint work with Niloufer Mackey, D. Steven Mackey,  
and Françoise Tisseur.



THE UNIVERSITY  
*of* MANCHESTER

# Group Background

Given nonsingular  $M$  and  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{C}$ ,

$$\langle x, y \rangle_M = \begin{cases} x^T M y, & \text{real or complex bilinear forms,} \\ x^* M y, & \text{sesquilinear forms.} \end{cases}$$

Define **automorphism group**

$$\mathbb{G} = \{ A \in \mathbb{K}^{n \times n} : \langle Ax, Ay \rangle_M = \langle x, y \rangle_M, \forall x, y \in \mathbb{K}^n \}.$$

Recall **adjoint**  $A^*$  of  $A \in \mathbb{K}^{n \times n}$  wrt  $\langle \cdot, \cdot \rangle_M$  defined by

$$\langle Ax, y \rangle_M = \langle x, A^* y \rangle_M \quad \forall x, y \in \mathbb{K}^{n \times n}.$$

Can show:  $A^* = \begin{cases} M^{-1} A^T M, & \text{for bilinear forms,} \\ M^{-1} A^* M, & \text{for sesquilinear forms.} \end{cases}$

$$\mathbb{G} = \{ A \in \mathbb{K}^{n \times n} : A^* = A^{-1} \}.$$

# Some Automorphism Groups

Space	$M$	$A^*$	Automorphism group, $\mathbb{G}$
Groups corresponding to a bilinear form			
$\mathbb{R}^n$	$I$	$A^T$	Real orthogonals
$\mathbb{C}^n$	$I$	$A^T$	Complex orthogonals
$\mathbb{R}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q} A^T \Sigma_{p,q}$	Pseudo-orthogonals
$\mathbb{R}^n$	$R$	$RA^T R$	Real perplectics
$\mathbb{R}^{2n}$	$J$	$-JA^T J$	Real symplectics
$\mathbb{C}^{2n}$	$J$	$-JA^T J$	Complex symplectics
Groups corresponding to a sesquilinear form			
$\mathbb{C}^n$	$I$	$A^*$	Unitaries
$\mathbb{C}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q} A^* \Sigma_{p,q}$	Pseudo-unitaries
$\mathbb{C}^{2n}$	$J$	$-JA^* J$	Conjugate symplectics

$$R = \begin{bmatrix} & & & 1 \\ & \cdot & \cdot & \\ & \cdot & \cdot & \\ 1 & & & \end{bmatrix}, \quad J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}, \quad \Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$$

# Questions

Recall **polar decomposition** of  $A \in \mathbb{C}^{n \times n}$ :

$$A = UH, \quad U^*U = I, \quad H = H^* \geq 0.$$

If  $A \in \mathbb{G}$

- ▶ When do its polar factors lie in the group?
- ▶ How can we exploit any group structure when computing  $U$  and  $H$ ?

# Structure of Polar Factors

Let  $\mathcal{U}$  denote set of groups for which  $M$  unitary.

Mackey, Mackey & Tisseur, 2003:

**Theorem 1** *Let  $\mathbb{G} \in \mathcal{U}$  and  $A \in \mathbb{G}$ . Then in  $A = UH$  the polar factors  $U$  and  $H$  belong to  $\mathbb{G}$ .*

# Structure of Polar Factors

Let  $\mathcal{U}$  denote set of groups for which  $M$  unitary.

Mackey, Mackey & Tisseur, 2003:

**Theorem 1** Let  $\mathbb{G} \in \mathcal{U}$  and  $A \in \mathbb{G}$ . Then in  $A = UH$  the polar factors  $U$  and  $H$  belong to  $\mathbb{G}$ .

**Theorem 2** Let  $\mathbb{G} \in \mathcal{U}$  and  $A \in \mathbb{G}$ . The singular values of  $A$  occur in reciprocal pairs  $\sigma$  and  $1/\sigma$ , with the same multiplicity.

# Structure-Preserving Iterations

**Theorem 3** Consider

$$Z_{k+1} = Z_k P_{mm}(I - Z_k^* Z_k) Q_{mm}(I - Z_k^* Z_k)^{-1}, \quad Z_0 = A,$$

where  $P_{mm}(t)/Q_{mm}(t)$  is the  $[m/m]$  Padé approximant to  $(1 - t)^{-1/2}$  and  $m \geq 1$ . If  $\mathbb{G} \in \mathcal{U}$  and  $A \in \mathbb{G}$  then

- $Z_k \in \mathbb{G}$  for all  $k$ ,
- $Z_k \rightarrow U$  at order  $2m + 1$ .

# Structure-Preserving Iterations

**Theorem 3** Consider

$$Z_{k+1} = Z_k P_{mm}(I - Z_k^* Z_k) Q_{mm}(I - Z_k^* Z_k)^{-1}, \quad Z_0 = A,$$

where  $P_{mm}(t)/Q_{mm}(t)$  is the  $[m/m]$  **Padé approximant** to  $(1 - t)^{-1/2}$  and  $m \geq 1$ . If  $\mathbb{G} \in \mathcal{U}$  and  $A \in \mathbb{G}$  then

- $Z_k \in \mathbb{G}$  for all  $k$ ,
- $Z_k \rightarrow U$  at order  $2m + 1$ .

Iterations  $z_{k+1} = f(z_k)$ :

$m$	$f(x)$
1	$\frac{x(3 + x^2)}{1 + 3x^2}$ <b>Halley</b>
2	$\frac{x(5 + 10x^2 + x^4)}{1 + 10x^2 + 5x^4}$



# Iterations (all with $X_0 = A$ )

Cubic (structure-preserving):

$$X_{k+1} = \frac{1}{3}X_k[I + 8(I + 3X_k^*X_k)^{-1}].$$

Quintic (structure-preserving):

$$x_{k+1} = x_k \left[ \frac{1}{5} + \frac{8}{5x_k^2 + 7 - \frac{16}{5x_k^2 + 3}} \right].$$

Scaled Newton iteration (not structure-preserving):

$$X_{k+1} = \frac{1}{2} \left[ \gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-*} \right], \quad \gamma^{(k)} = \left( \frac{\|X_k^{-1}\|_F}{\|X_k\|_F} \right)^{1/2}.$$

$$X_0 \in \mathbb{G} \in \mathfrak{u} \quad \Rightarrow \quad \gamma^{(0)} = 1.$$

# Experiment

Random symplectic  $A \in \mathbb{R}^{12 \times 12}$ ,  $\|A\|_2 = 310 = \|A^{-1}\|_2$ .

$$\mu_{\mathbb{O}}(A) = \frac{\|A^*A - I\|_2}{\|A\|_2^2}, \quad \mu_{\mathbb{G}}(A) = \frac{\|A^*A - I\|_2}{\|A\|_2^2}.$$

$k$	Newton (scaled)		Cubic	
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$
0	1.0e+0	7.0e-18	1.0e+0	7.0e-18
1	1.0e+0	1.0e+0	1.0e+0	8.9e-17
2	8.6e-01	8.6e-01	1.0e+0	8.1e-16
3	2.0e-01	2.0e-01	9.9e-01	6.3e-15
4	3.2e-03	3.2e-03	9.4e-01	5.0e-14
5	9.0e-07	9.0e-07	5.7e-01	2.8e-13
6	6.0e-14	1.3e-13	3.6e-02	5.2e-13
7	4.3e-16	1.1e-13	3.2e-06	5.3e-13
8			3.8e-16	5.3e-13

# Newton Behaviour

**Theorem 4** Let  $\mathbb{G} \in \mathfrak{U}$ ,  $A \in \mathbb{G}$ , and  $X_k$  be the Newton iterates, either unscaled or with Frobenius scaling. Then

$$X_k^\star = X_k^*, \quad k \geq 1. \text{ Moreover,}$$

$$MX_k = X_kM, \quad \text{real bilinear, complex sesquilinear forms,}$$

$$MX_k = \overline{X_k}M, \quad \text{complex bilinear forms.}$$

Implications:

★ Tethering.

★ Structure in  $X_k$ :

A pseudo-orthogonal  $\Rightarrow X_k$  block-diagonal,

$$\text{A symplectic} \Rightarrow X_k = \begin{bmatrix} E_k & F_k \\ -F_k & E_k \end{bmatrix}.$$

# Convergence Tests

Padé iteration function  $f_{mm}$  satisfies

$$f_{mm}(\sigma^{-1}) = f_{mm}(\sigma)^{-1}, \quad f_{mm}(1) = 1,$$

$$1 < \sigma \Rightarrow 1 < f_{mm}(\sigma) < \sigma,$$

$$1 \leq \mu < \sigma \Rightarrow f_{mm}(\mu) < f_{mm}(\sigma).$$

Can show that for  $A \in \mathbb{G}$  and  $\mathbb{G} \in \mathfrak{U}$ :

$$\|U - Z_k\|_2 = f_{mm}^{(k)}(\sigma_1) - 1.$$

Moreover,  $\|Z_k\|_F$  decreases monotonically. Stop when

$$\frac{\|\hat{Z}_{k+1}\|_F}{\|\hat{Z}_k\|_F} \geq 1 - \delta.$$

Scaled Newton: latter test applicable for  $k \geq 1$ .

# Structured Iterations Versus Scaled Newton

- ★ Newton has slightly better observed numerical stability.
- ★ Newton usually requires the fewest flops.
- ★ Convergence prediction possible with structured iterations.
- ★ Which is best depends on matrix  $A$ , group  $\mathbb{G}$ , and user's accuracy requirements.

Have analogous results & iterations  
for matrix sign function, with  
*no restrictions on  $\mathbb{G}$ .*

# Generalized Polar Decomposition

**Theorem 5** *Let  $\mathbb{G}$  be a group for which  $(A^\star)^\star = A$  for all  $A \in \mathbb{K}^{n \times n}$ . For any  $A \in \mathbb{K}^{n \times n}$  such that  $A^\star A$  has no eigenvalues on  $\mathbb{R}^-$ ,  $A$  has a unique decomposition*

$$A = WS,$$

where

$$\begin{aligned} W &\in \mathbb{G} && (\text{i.e., } W^\star = W^{-1}), \\ S^\star &= S, \end{aligned}$$

and  $\lambda(S)$  lies in the open right half-plane (i.e.,  $\text{sign}(S) = I$ ).

# Matrix Sign Relations

For nonsingular  $A \in \mathbb{C}^{n \times n}$  (Byers, 1984):

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix},$$

where  $A = UH$  is the polar decomposition.

# Matrix Sign Relations

For nonsingular  $A \in \mathbb{C}^{n \times n}$  (Byers, 1984):

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix},$$

where  $A = UH$  is the polar decomposition.

For  $A \in \mathbb{C}^{n \times n}$  with no eigenvalues on  $\mathbb{R}^-$  (H, 1997):

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}.$$



# More General Matrix Sign Relation

**Theorem 6** *Let  $A, B \in \mathbb{C}^{n \times n}$  and suppose  $AB$  has no eigenvalues on  $\mathbb{R}^-$ . Then*

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix},$$

*where  $C = A(BA)^{-1/2}$ .*

**Proof.** Use  $\text{sign}(P) = P(P^2)^{-1/2}$ .  $\square$

# More General Matrix Sign Relation

**Theorem 6** *Let  $A, B \in \mathbb{C}^{n \times n}$  and suppose  $AB$  has no eigenvalues on  $\mathbb{R}^-$ . Then*

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix},$$

where  $C = A(BA)^{-1/2}$ .

**Proof.** Use  $\text{sign}(P) = P(P^2)^{-1/2}$ .  $\square$

**Corollary 1** *Let  $A \in \mathbb{K}^{n \times n}$  have a generalized polar decomposition  $A = WS$ . Then*

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & W \\ W^* & 0 \end{bmatrix}.$$

# Puzzle

Proof of previous theorem shows that

$$A(BA)^{-1/2} = [B(AB)^{-1/2}]^{-1} = (AB)^{1/2}B^{-1}.$$

Why do we have equality?

# Puzzle

Proof of previous theorem shows that

$$A(BA)^{-1/2} = [B(AB)^{-1/2}]^{-1} = (AB)^{1/2}B^{-1}.$$

**Lemma 1** *Let  $A, B \in \mathbb{C}^{n \times n}$  and  $f(AB)$  and  $f(BA)$  be defined. Then*

$$Af(BA) = f(AB)A.$$

# Puzzle

Proof of previous theorem shows that

$$A(BA)^{-1/2} = [B(AB)^{-1/2}]^{-1} = (AB)^{1/2}B^{-1}.$$

**Lemma 1** *Let  $A, B \in \mathbb{C}^{n \times n}$  and  $f(AB)$  and  $f(BA)$  be defined. Then*

$$Af(BA) = f(AB)A.$$

**Proof.** There is a *single* polynomial  $p$  such that  $f(AB) = p(AB)$  and  $f(BA) = p(BA)$  (e.g.,  $p$  interpolating to  $f$  on product of minimum polys of  $AB$  and  $BA$ ). Hence

$$Af(BA) = Ap(BA) = p(AB)A = f(AB)A. \quad \square$$

# Puzzle

Proof of previous theorem shows that

$$A(BA)^{-1/2} = [B(AB)^{-1/2}]^{-1} = (AB)^{1/2}B^{-1}.$$

**Lemma 1** *Let  $A, B \in \mathbb{C}^{n \times n}$  and  $f(AB)$  and  $f(BA)$  be defined. Then*

$$Af(BA) = f(AB)A.$$

Now

$$\begin{aligned} A(BA)^{-1/2} \cdot B(AB)^{-1/2} &= (AB)^{-1/2}A \cdot B(AB)^{-1/2} \\ &= (AB)^{-1/2}AB(AB)^{-1/2} \\ &= I. \end{aligned}$$